Identifying Patients for Clinical Studies from Electronic Health Records: TREC 2012 Medical Records Track at OHSU

Steven Bedrick[1], Tracy Edinger[2], Aaron Cohen[2], William Hersh[2]
[1]Center for Spoken Language Understanding
[2]Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University, Portland, OR, USA

The goal of the TREC 2012 Medical Records Track was to search medical record documents to identify patients as possible candidates for clinical studies based on diagnosis, age, and other attributes. For TREC 2012, the Oregon Health & Science University (OHSU) group experimented with both manual and automated techniques. We used a derivative of Lucene to build an interactive retrieval system that can process queries in one of two ways. Users can manually specify Boolean queries whose terms may include words as well as ICD-9 codes. Alternatively, the system features an automated query parser that transforms free-text queries into structured Boolean queries. The query parser is built on top of MetaMap and the UMLS Metathesaurus. We submitted both automatic runs (which relied solely on the automated query parser) as well as manual runs consisting of queries built by an expert clinician. Overall, our automated query parser performed below the mean of other groups, although there were individual topics for which it performed very well. This irregular performance was in part due to our parser's tendency to over-specify queries, leading to reduced recall. There were, however, several topics for which our parser performed very well, suggesting that our fundamental approach has merit. In contrast, our manual runs performed very well, scoring second-best among official manual runs. With further modification of the manual queries, we were able to achieve even better performance. Query of electronic health records for the use case of identifying patients as candidates for clinical studies still requires manual query development, at least until better automated methods can be developed that outperform them.

Introduction

The task of the TREC 2012 Medical Records Track (TRECMed), similar to TREC 2011, consisted of searching electronic health record (EHR) documents in order to identify patients matching a set of clinical criteria, a use case that might be part of the preparation of a quality report or to develop a cohort for a clinical trial [1]. The task's various topics each represented a different case definition, with the topics varying widely in terms of detail and linguistic complexity. This use case is one of a larger group that represent the "secondary use" of data in EHRs [2] that facilitate clinical research, quality improvement, and other aspects of a health care system that can "learn" from its data and outcomes [3]. It is made possible by the large US government investment in EHR adoption that has occurred since 2009 [4].

The corpus for TRECMed 2012 was identical to TRECMed 2011 and consisted of a set of 93,552 patient encounter files extracted from an EHR system. Each encounter file represented a note entered by a clinician or a report in the course of caring for a patient. Each note or report was categorized by type (e.g., History & Physical, Surgical Pathology Report, Radiology Report) or in some cases sub-type (e.g., Angiography). Certain items, such as name, age, and address, were de-identified.

The encounter files were each associated with one of 17,265 unique patient visits to the hospital or emergency department. Most visits (≈70%) included five or fewer encounters; virtually all (≈97%)

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Identifying Patients for Clinical Studies from Electronic Health Records: TREC 2012 Medical Records Track at OHSU	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Oregon Health & Science University,Center for Spoken Language Understanding,Portland,OR**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**18**	

included less than 20. The maximum number of encounters comprising any visit was 415. Each encounter within a visit shared a short (truncated to 40 characters) “chief complaint” as well as a single “admission” ICD-9 code and a set of “discharge” ICD-9 codes. The number of discharge ICD-9 codes varied widely from visit to visit; the median number of codes per visit was 5, while the maximum was 25. (Note, however, that the number of visits with 25 discharge codes was more than four times greater than the number of visits with 24 codes; for this reason, we suspect that the “true” maximum number of codes per visit may have been higher in the original EHR that gave rise to the corpus, and that the apparent limit of 25 codes per visit is simply an artifact of the export process, much in the same way as the chief complaint field’s length was truncated to 40 characters.)

Patients could not be linked across visits, i.e., be identified as having more than one visit, due to the de-identification process. As such, for the purposes of this task, the “unit of retrieval” was the visit rather than the patient, meaning that the participating systems were to produce a set of matching visits for each topic. Visits could not be tied to individual patients and therefore visit was used as a surrogate for an individual patient meeting the given clinical criteria.

The Oregon Health & Science University (OHSU) team took a manual, interactive approach to the task, and focused on the construction of a search interface that would allow its users to rapidly formulate queries, review their results, and iterate. Using our system, users could search through the corpus by any of the various fields (chief complaint, report text, etc.) using a robust search syntax, and could also include ICD-9 codes in their queries. This allowed for the easy construction of queries representing complex Boolean criteria.

Methods

In TRECMed 2012, we continued work on the retrieval system that we used during TRECMed 2011 [5]. In particular, we laid the foundation of a "natural-language" query parser that is intended to take free-text topic descriptions and translate them into our system's query syntax. We will provide a brief overview of last year's system, and then proceed to describe the operation of our query parser.
Our existing retrieval system is based on a port of Apache Lucene to the Ruby programming language [6]. As such, it supports the standard Lucene search syntax. We have extended the out-of-the-box system by adding modules to correctly index ICD-9 codes, and to allow "wildcard" searching of said codes. Our system provides a flexible web interface to the system, and allows users to use several different methods to search. Users can also easily browse through the document collection by various fields (ICD-9 codes, etc.) The system can produce output in the form of a standard page-of-links, or it can directly output trec_eval-formatted run files. Our paper from last year goes into detail about the system's operating characteristics [5].

One major shortcoming of our system from last year is its poor ability to handle free-text queries. For anything other than trivial keyword searches, users must learn the Lucene search syntax, which---powerful though it is--- is a complex task for a non-technical user. This year, we wanted to improve on this situation, and provide users a way to interact with the system in a more natural way.

Our query parser uses the MetaMap tool from the National Library of Medicine (NLM) to analyze the text of user queries and to identify any medical concepts that are present [7]. We then use the UMLS to perform query expansion, as well as to identify candidate ICD-9 codes for inclusion in the final query. This approach is simplistic, but serves as a starting point for more sophisticated methods going forward.

The parser works as follows. The system passes raw queries to the May, 2012 MetaMap release, which performs two relevant actions: it segments the query into discrete phrases, and maps each phrase to some number of Unified Medical Language System (UMLS) Metathesaurus concepts [6 ***not Endnote]. We treat each phrase as a component of the final query; the version of the parser described by this paper makes the assumption that the query components should be linked using a Boolean "AND". This is a significant limitation in that it mishandles queries that include qualifying text ("patients with symptom AAA who were not given drug BBB"); future versions will make use of more sophisticated phrase chunking and analysis methods.

The parser analyzes each component phrase in turn. For each phrase, MetaMap provides a list of possible "mappings" from that phrase to a concept in the UMLS Metathesaurus [8]. Each mapping consists of a UMLS concept unique identifier (CUI) as well as a score indicating MetaMap's confidence in that mapping. Our parser only uses mappings whose scores fall above a configurable confidence threshold; for the experiments described in this paper, we sought to avoid overly "noisy" mappings and so used a high threshold of 800 (on a score of 0-1,000).[1]

Once we have determined that a given candidate mapping to a phrase is strong enough to consider, we then check it against a short list (23 items) of "stop-CUIs" - CUIs that we have manually decided to exclude. There are several UMLS Metathesaurus concepts that occur in the TRECMed corpus with sufficient frequency to render them at best uninformative and at worst harmful to the final query. Some examples include C0030705 ("Patients"), C0332293 ("Treated with"), C1514756 ("Receive"), and so on. If a candidate mapping is included in this list, we do not include it in our final query. We constructed this list by experimenting with the TRECMed 2011 topics and observing which MetaMap mappings resulted in particularly un-helpful concepts. After a candidate mapping has passed the stop-CUI check, our parser examines its "semantic type." We maintain a "whitelist" of semantic types that we are interested in including (see Table 1); mappings that are not of one of these types are removed from consideration.

For TRECMed 2012, we chose to restrict ourselves to terms that exist in the MeSH vocabulary. Our reasoning was initially that we would be able to take advantage of MeSH's hierarchical nature to include hyper/hyponyms in our expanded queries. While we ultimately did not find that particular approach to be an effective one, we found two other aspects of MeSH to be particularly useful for our purposes: first, MeSH contained reasonably broad coverage over the types of concepts contained in our training set of topics; and second, that the curators of MeSH have included a rich and diverse set of synonymous entry terms for most of the vocabulary's contents. As such, after ensuring that mapped concepts were not in the list of stop-CUIs, our next check was to attempt to map the CUI to a MeSH entry.[2] If no such entry existed, the mapped concept was not in MeSH, and so our parser ignores it and goes on to the next candidate mapping.

[1] For the sake of simplicity, in our discussion we have reversed the sign of the threshold. In reality, MetaMap's scoring uses negative numbers wherein "lower is better". Our actual threshold was therefore "-800."

[2] See this paper's "Discussion" section for a description of a version of our parser that experimented with incorporating non-MeSH mappings.

Table 1 - Semantic type whitelist.

Disease or Syndrome
Pharmacologic Substance
Neoplastic Process
Therapeutic or Preventative Procedure
Sign or Symptom
Health Care Activity
Medical Device
Body Location or Region
Mental or Behavioral Dysfunction

At this point, the remaining component phrases have been translated into sets of CUIs that are known to have corresponding entries in MeSH. The next step for our parser is to expand each of these to its set of entry terms, which our parser can accomplish using either the UMLS Metathesaurus or the standard XML distribution of MeSH. Each entry term ends up linked together with "OR" clauses in the final query.[3] For example, for topic 165 ("Patients who have gluten intolerance or celiac disease"), "gluten intolerance" is mapped to CUI C0007570, which has several different MeSH entry terms: "celiac disease", "gluten enteropathy", "gluten-sensitive enteropathy", "nontropical sprue", etc. The resulting query fragment for that part of the topic would look something like "'celiac disease' OR 'gluten enteropathy' OR ...".

During TRECMed 2011, we found that it was often useful to include ICD-9 codes in our queries. Therefore, in addition to resolving entry terms for mapped concepts, our parser can also attempt to look up relevant ICD-9 codes.[4] It does this using the UMLS, and can include them in the final query in one of two ways. The first (and standard) way is to simply include the ICD-9 code in the same query fragment as its originating CUI. So, in the "celiac disease" example, the final query fragment would consist of the various entry terms followed by " OR icd:579.0"[5]. It is also possible to configure the parser such that, instead of including the ICD-9 codes "in-line" with the entry terms, it will append them to the end of the query as a completely separate "OR" clause. For our submitted runs, we used the first ("in-line") method.

After the mapping and expansion has taken place, our query parser assembles the phrase/entry-term-list pairs into a final query. As previously mentioned, each phrase ends up being represented as a list of entry terms "OR"-ed together; we are also able to include the original text that gave rise to the mapping in the first place in this list. Each phrase's resulting sub-query is grouped with parentheses, and then linked together using the Boolean "AND" operator.

For an example of what this process results in, consider topic 177 ("Patients treated for depression after myocardial infarction"). Depending on ICD code placement, our parser can construct the final query in two different ways:

---

[3] We also perform very rudimentary stop-word removal within entry terms, involving words such as "the", "and", etc.

[4] This aspect of our query parser's behavior is fully configurable, as are nearly all other aspects of its operation.

[5] The "icd:" prefix tells the search system to only consider terms found in the "icd" index field; this, however, is just an example- in reality, our system makes uses of a variety of index fields, and all terms in the queries generated by our parser include field specifiers. For the sake of clarity we have omitted (or shortened, in the case of the ICD-9 example) the field names from the sample queries in this paper.

1. ("depressive disorders" OR "mental depression" OR ... OR icd:311) AND ("myocardial infarction" OR "myocardial infarct" OR ... OR icd:410.9)
2. (("depressive disorders" OR "mental depression" OR ...) AND ("myocardial infarction" OR "myocardial infarct" OR ...)) OR (icd:311 AND icd:410.9)

We submitted three official automatic runs for TRECMed 2012. The first, *OHSUCombET*, used the original query terms, and both the the MetaMap "preferred" term as well as the additional entry terms. The second, *OHSUCombICD*, used the original query terms, the MetaMap "preferred" term, and the relevant ICD-9 codes. The last, *OHSUCEtICD*, used all three: the mapped preferred term, sibling entry terms, and ICD-9 codes.

In addition to the automatic runs, we also submitted a single manual official run, *ohsuManBool*. Boolean queries for this run were constructed with ICD-9 codes and phrases. For each condition in the topic, all relevant ICD-9 codes were listed and joined with *OR*. Possible names or descriptions of each condition were also included in this list, using wildcards to capture suffix variants. Similarly, if a drug was part of the topic, all names and variations for that drug were joined with *OR*.

In some cases, the topic specified details not related to a condition or drug. For example, some topics specified treatment in the emergency room, and others indicated a particular age of patient. These requirements were included in the query by listing phrases used to refer to them. To identify patients seen in the emergency room, we searched for *emergency room OR emergency department.*

The de-identification process rendered all age references in the format ***AGE[in 30s]*. For topics that specified an age or age range, our query included a list of relevant ages separated by *OR*. Several topics specified ages that could not be accurately represented because of the de-identified age ranges. For example, children (age less than 18) and adults over 65 could not be precisely identified in queries, and the queries captured additional visits or missed some visits. After each segment of the query was constructed, all segments were joined with *AND*.

Table 2 - Selected performance metrics for OHSU runs.

Run	infAP	infNDCG	P @ 10
OHSUCEtICD	0.124	0.291	0.319
OHSUCombET	0.110	0.278	0.321
OHSUCombICD	0.131	0.269	0.328
ohsuManBool	0.250	0.526	0.611
All Runs	0.169	0.424	0.470
All Automatic Runs	0.170	0.424	0.470
All Manual Runs	0.200	0.465	0.551
Best Automatic Run	0.286	0.578	0.592
Best Manual Run	0.366	0.680	0.749

Results

OHSU submitted a total of four runs, one of which was manual and the other three of which were automatic. The median results of our runs and then the median of all runs, all automatic runs, and all manual runs are shown in Table 2.

Manual Run

Overall, our manual run performed quite well. There were twelve topics for which it provided the best results (in terms of infNDCG), and 21 for which it performed above the median. Looking at early precision (P@10), our manual run provided the best results for sixteen topics and exceeded the median P@10 score for 21 topics.[6] As also noted in **Error! Reference source not found.**, our manual run outperformed the median in terms of infAP, infNDCG, and P@10. Of particular note was topic 179 ("Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression"), for which our manual run outperformed the median infNDCG score by more than a factor of ten (0.72 vs. 0.06).

While our manual run's performance was quite good in general, there were several problematic topics. Our manual run did not retrieve any relevant results for topic 172 ("Patients with peripheral neuropathy and edema"), which was surprising given that this does not appear to have been a particularly difficult topic for other systems (given the topic's median manual infNDCG score of 0.317- not particularly high, but far from the lowest). The concept of edema proved to be quite problematic. In developing the query, we specified the ICD-9 code or the term "edema," and the codes or term for "peripheral neuropathy." Inspection of the resulting visits revealed that most were captured because of documentation of a *lack* of edema. We limited our search to the ICD-9 code for edema, but clearly that was over-restrictive.

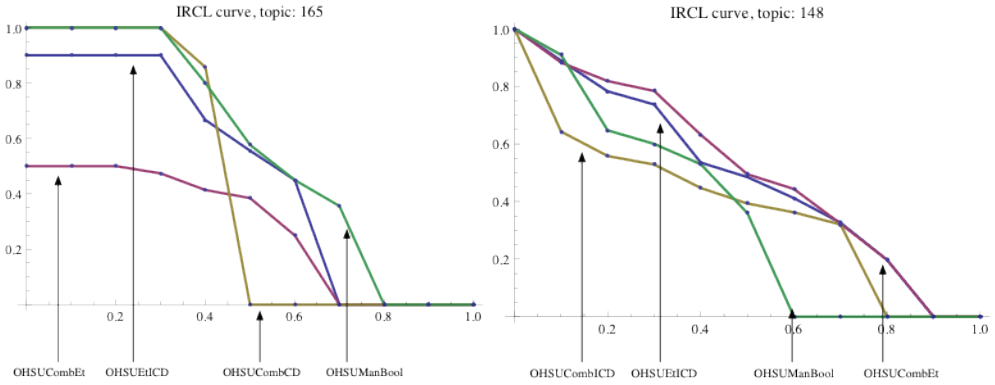
Another problematic topic was 145 ("Patients with lupus nephritis and thrombotic thrombocytopenic purpura"), for which we achieve an infNDCG of 0.267- well below the median of 0.6722, and the lowest-scoring run for this topic in terms of infNDCG. In this case, the low performance was the result of Boolean over-specification as well as possibly an over-dependence on using ICD-9 codes, as only a very small number of documents in the collection include both the ICD-9 codes for lupus *and* for thrombotic thrombocytopenic purpura, so a simplistic approach to integrating ICD-9 codes into the query will retrieve very few results.

Automatic Runs

Our automatic runs did not perform particularly well, overall (see Table 2). Averaged over all topics, all three runs performed well below the median in terms of infAP, infNDCG, and P@10. There were also numerous (15) topics for which none of the automatic runs were able to retrieve *any* relevant results, and another two for which that was the case for at least one automatic run.

[6] There were 35 topics for which our manual run *met or exceeded* the median level.

Figure 1 Recall-precision curves for topics 165 and 148.



There were, however, some topics where this approach performed well. For example, our automatic runs that used ICD-9 codes as components of the query achieved the best infNDCG score on topic 154 ("Patients with Primary Open Angle Glaucoma (POAG)"), at 0.8895. There are several semantically-similar UMLS Metathesaurus concepts relating to various forms of glaucoma, and MetaMap's output for that topic included many of them. This, in turn, meant that our query included several different glaucoma-related ICD-9 codes- in addition to 365.1, for open-angle glaucoma, the final query included the "NOS" version of 365.1 (365.10), 365.13 (pigmentary glaucoma), and both 365 and 365.9 (glaucoma and "glaucoma (NOS)"). This meant that our query retrieved essentially every glaucoma-related visit, but because it did also contain useful query terms such as "open" and "angle," our system was able to achieve a reasonable result ranking, and our ICD-containing automatic runs' precision scores were quite high for this topic.

There were also several topics for which one or another of our automatic runs outperformed the median, often by a significant amount. For example, our "entry-terms" automatic run's infNDCG score of 0.545 for topic 185 ("Patients who develop thrombocytopenia in pregnancy") was more than twice the median score for that topic. Interestingly, our runs that included ICD-9 codes performed somewhat worse than did this run, although both still came in over the median score.

Additional Runs

In addition to the official runs described above, we prepared an additional set of runs after the TREC conference, based on making minor modifications to our query parser and manual queries. The modification involved changing the vocabularies available to the parser for query expansion. As noted earlier, our system in its initial form relied entirely on MeSH headings. We wanted to determine whether including synonymous terms from non-MeSH vocabularies would have any effect on system performance; in other words, expanding the list of possible query expansions from MeSH entry terms to UMLS sibling-terms (term unique identifiers or LUIs, in UMLS parlance). This relatively simple change had a significant impact on our system's performance. Unsurprisingly, including terms from additional vocabularies had a dramatic effect on the length of the queries: the median query length for our submitted OHSUComboEt run was 503 characters, whereas an equivalent run generated using the modified query parser had a median length of 2,850 characters. This increase has an obvious cost in terms of comprehensibility as well as system processing time.

However, the longer queries contained a much more diverse set of terms, which had a generally positive effect on the retrieval performance of the system. We created two modified runs, one with and one without ICD codes ("plus_mesh_icd" and "plus_mesh", comparable to OHSUCEtICD and OHSUCombEt, respectively). plus_mesh_icd achieved an overall infNDCG score of 0.2916, which was higher than our best-performing officially-submitted run (plus_mesh achieved an overall infNDCG score of 0.2791).

Including terms from additional vocabularies had a strong effect on certain topics, in both positive and negative ways. Topic 161 ("Patients with adult respiratory distress syndrome") saw a dramatic improvement in infNDCG between plus_mesh and our official runs ($\approx+\Delta 0.43$ between plus_mesh and OHSUCombEt), and several others saw more modest improvements. However, some topics suffered decreased performance, possibly due to an increase in the noise level of the results.

We also performed one additional manual run. We obtained the queries developed by the NLM group, whose manual queries achieved the highest performance of all official runs in TRECMed 2012 [9].[7] We compared our queries to theirs and added any new terms from their queries. When applicable, we included their Boolean logic. Their search system was very sophisticated, allowing the capability to weight terms and search specific sections of the record, and their queries reflected this. In a few cases, the terms used in the queries were so embedded in the search constructs that we could not add them to our queries.

In most cases, we obtained modest increases in performance. In two queries, we found a significant change in the number of relevant documents retrieved. In topic 173 ("Patients over 65 who had Pneumonia Vaccination Status presently or previously"), we added the restriction that the report not document *"no pneumonia vaccine."* This lowered the number of relevant documents retrieved by 35, so we decided not to include this phrase. A sample of documents revealed that the vaccination may have been completed after documenting its absence, so this phrase eliminated relevant documents. Our original query for topic 177 ("Patients treated for depression after myocardial infarction"), looked for a list of ICD-9 codes for myocardial infarction and a mention of an antidepressant. Adding the phrase *"myocardial infarction"* to the list of ICD-9 codes increased the number of relevant documents retrieved by 35.

Overall, our new scores were improved over our original scores. Our infAP and infNDCG increased to 0.278 and 0.565 respectively. In spite of this improvement, we still did not reach the scores achieved by the NLM group. Their high scores resulted from their ability to construct highly specific weighted queries, not solely as a function of selecting the best search terms. The summarized results of the additional automatic and manual runs are listed in Table 3. Our final manual queries are listed in Appendix 1.

Table 3 - Additional post-TREC conference runs.

Run	infAP	infNDCG	P @ 10
plus_mesh_icd	0.127	0.292	0.313
plus_mesh	0.117	0.276	0.296
OHSUManualRun3	0.278	0.565	0.642

[7] We thank Dina Demner-Fushman and her NLM group for graciously sharing their queries with us.

Discussion

In general, our manual runs outperformed the automatic runs. However, there were several exceptions to this tendency. As previously mentioned, our official manual run did not retrieve any relevant documents for topic 172 ("Patients with peripheral neuropathy and edema"). Our automatic runs did not have this problem, and in fact performed reasonably well - our entry-term-only run beat the median for this topic with an infNDCG score of 0.3390. Our automatic runs that made use of ICD-9 data, however, did not perform quite as well, presumably due to the same ICD-9 issues as bedeviled our manual run. This suggests that our automatic query expansion approach was relatively effective as compared to a stricter, "concept-based" approach.

This difference in performance between automatic runs was far from unusual. The first chart in **Error! Reference source not found.** shows the precision/recall graph for all four of our runs for topic 165 ("Patients who have gluten intolerance or celiac disease"), chosen here because it illustrates a common pattern that we observed in our results. For some topics, such as this one, relying solely on entry terms was insufficient, and the query results benefitted from the inclusion of ICD codes (as evidenced by the higher early precision achieved by OHSUCombICD). However, note that the results for OHSUCombICD (which did not include entry term expansion of the query) suffered a sharper drop-off in precision than did OHSUEtICD (which included both entry terms as well as ICD codes).

There were also cases where including ICD codes hurt query performance; the second chart in **Error! Reference source not found.** shows the precision/recall graph for topic 148 ("Patients acutely treated for migraine in the emergency department"). Note that the run that did not include query term expansion performed substantially worse than the others. On balance, then, we found that the combination of ICD code inclusion as well as query term expansion generally performed the most reliably across topics. Future work may involve attempting to automatically recognize queries for which one approach or the other would be optimal.

One major issue with our automatic runs was the relatively large number of runs for which we returned no relevant results. Often, this was the result of Boolean over-specification as a result of our parser choosing to include too many query clauses or improperly processing the clauses that it included. An instructive example is topic 141, “Adult inpatients with Alzheimer's disease admitted from nursing homes with pressure ulcers.” Our query parser produced a query with three clauses- one dealing with Alzheimer’s disease, another with nursing homes, and a final clause concerning pressure ulcers. However, because the semantic type of the “nursing home” concept was not in our whitelist, no MetaMap-derived data (and therefore no entry terms or ICD codes) were included. In such a case (where there is a high-scoring mapped concept that is not on the blacklist of concepts to avoid, but is not of a whitelisted semantic type), our parser includes the original query text that gave rise to the mapping under the assumption that it contains relevant information. In this case, the phrase that was included was “from nursing homes,” which was a somewhat longer and grammar-laden fragment than we had been anticipating when we designed this behavior into our parser.

The other two parts of the query (Alzheimer’s disease and pressure ulcers) were both processed reasonably well by the query parser, and the three components were linked with a Boolean “AND” operation. Since (apparently) no documents in the collection contained the phrase “from nursing homes”, our retrieval system returned no results. This sort of problem was far from uncommon; future work on this system will focus on better handling of corner cases such as these.

Clearly, our approach to automatic query parsing has both advantages and disadvantages. On the one hand, it is simple to understand and is capable of producing reasonable-looking queries. On the other

hand, it is brittle, and heavily dependent on both MetaMap's mapping ability as well as the linguistic vagaries of MeSH terms. Our additional automatic runs attempted to address this issue by including vocabulary from sources other than MeSH, and while we saw an overall increase in performance, we also saw several notable decreases in performance.

The automatically generated queries themselves also became longer and more complex, and we suspect that this increased complexity sometimes introduced "stray" (i.e., unnecessary and potentially incorrect) terms into the query. Overall, we feel that increasing the number of vocabularies in use is generally beneficial, but there were enough exceptions that we are unwilling to categorically recommend it. Future work will investigate more effective means of determining whether or not a given query expansion term should be included in the final query.

In some ways, our approach is similar to that described by Koopman, et al. from the 2011 TRECMed track [7]. There are two salient differences, however. First, Koopman et al. represent visits as vectors of MetaMap-provided concepts, and second, they map the MetaMap-provided UMLS CUIs to their equivalent SNOMED-CT concepts. They refer to their approach as "bag of concepts" retrieval in that both the query and the document are represented by sets of SNOMED concepts. Our approach, in contrast, uses MeSH rather than SNOMED,[8] and represents documents using a standard term vector model.

Conclusions & Future Work

Our manual Boolean query run achieved excellent results, showing that at least until the content of EHRs (as well as methods for indexing and searching them) is better characterized, human-constructed queries will continue to be required for the retrieval of medical records, particularly for the use case of identifying patients who are candidates for clinical studies. This suggests that one potentially valuable area of research in medical IR would be the development of improved tools to support clinicians in building and evaluating queries over clinical document repositories. In future work, we aim to improve the performance of our manual queries as well as use them as models for new methods. We will also continue to iterate our automated query parser, and hope to make it more resilient to linguistic variation in topic descriptions.

References

1. Voorhees E and Hersh W. *Overview of the TREC 2012 Medical Records Track. The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)*. 2012. Gaithersburg, MD: National Institute for Standards and Technology.
2. Safran C, Bloomrosen M, Hammond WE, Labkoff SE, Markel-Fox S, Tang P, et al., *Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper.* Journal of the American Medical Informatics Association, 2007. 14: 1-9.
3. Friedman CP, Wong AK, and Blumenthal D, *Achieving a nationwide learning health system.* Science Translational Medicine, 2010. 2(57): 57cm29. http://stm.sciencemag.org/content/2/57/57cm29.full.

[8] There is, of course, no reason in principle why our approach would not work with SNOMED instead of MeSH. Both vocabularies are reasonably diverse in terms of their linguistic content, and so would presumably work similarly well for query expansion.

4. Blumenthal D, *Wiring the health system--origins and provisions of a new federal program.* New England Journal of Medicine, 2011. 365: 2323-2329.
5. Bedrick S, Ambert K, Cohen A, and Hersh W. *Identifying Patients for Clinical Studies from Electronic Health Records: TREC Medical Records Track at OHSU. The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*. 2011. Gaithersburg, MD: National Institute for Standards and Technology. http://trec.nist.gov/pubs/trec20/papers/OHSU.medical.update.pdf.
6. McCandless M, Hatcher E, and Gospodnetic O, *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. 2010, Greenwich, CT: Manning Publications.
7. Aronson AR and Lang FM, *An overview of MetaMap: historical perspective and recent advances.* Journal of the American Medical Informatics Association, 2010. 17: 229-236.
8. Bodenreider O, *The Unified Medical Language System (UMLS): integrating biomedical terminology.* Nucleic Acids Research, 2004. 32: D267-D270.
9. Demner-Fushman D, Abhyankar S, Jimeno-Yepes A, Loane R, Lang F, Mork JG, et al. *NLM at TREC 2012 Medical Records Track*. *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)*. 2012. Gaithersburg, MD: National Institute for Standards and Technology.

Appendix 1 - OHSU Manual Queries

From Final Run (OHSUManualRun3)

136 Children with dental caries

```
(discharge_icd_codes_txt:521* OR ((report_text:cavit* OR report_text:caries) AND
(report_text:tooth OR report_text:teeth OR report_text:dent*))) AND
(report_text:"birth-12" OR report_text:"**AGE[<> teens]") AND NOT
(report_text:"**AGE[<> 20s" OR report_text:"**AGE[<> 30s" OR report_text:"**AGE[<>
40s" OR report_text:"**AGE[<> 50s" OR report_text:"**AGE[<> 60s" OR
report_text:"**AGE[<> 70s" OR report_text:"**AGE[<> 80s" OR report_text:"**AGE[90+")
```

137 Patients with inflammatory disorders receiving TNF-inhibitor treatments

```
(report_text:"adalimumab|humira|infliximab|remicade|cerolizumab|cimzia|golimumab|simpo
ni|etancercept|enbrel|trental" OR report_text:"tnf-inhibitor" OR report_text:"tnf
inhibitor") AND
(discharge_icd_codes_txt:"286.53|517.8|601.1|601.9|135|595.1|696.0|696.1|701.0" OR
discharge_icd_codes_txt:274* OR discharge_icd_codes_txt:446* OR
discharge_icd_codes_txt:582* OR discharge_icd_codes_txt:583* OR
discharge_icd_codes_txt:556* OR discharge_icd_codes_txt:614* OR
discharge_icd_codes_txt:615* OR discharge_icd_codes_txt:616* OR
discharge_icd_codes_txt:710* OR discharge_icd_codes_txt:714* OR
discharge_icd_codes_txt:720* OR discharge_icd_codes_txt:728* OR
discharge_icd_codes_txt:558.4*)
```

138 Patients with acute tubular necrosis due to aminoglycoside antibiotics

```
(discharge_icd_codes_txt:584.5 OR report_text:"acute tubular necrosis") AND
report_text:"amikacin|apramycin|arbekacin|astromicin|bekanamycin|capreomycin|dibekacin
|dihydrostreptomycin|elsamitrucin|fosfomycin|g418|gentamicin|hygromycin|isepamicin|kan
amycin|kasugamycin|lividomycin|micronomicin|neamine|neomycin|netilmicin|paromomycin|ri
bostamycin|sisomicin|streptoduicin|streptomycin|tobramycin|verdamicin|aminoglycoside|n
etilmicin|spectinomycin" AND NOT report_text:"cholestin"
```

139 Patients who presented to the emergency room with an actual or suspected miscarriage

```
 (discharge_icd_codes_txt:640.* OR discharge_icd_codes_txt:641.* OR
discharge_icd_codes_txt:V22.2 OR discharge_icd_codes_txt:634.11 OR
discharge_icd_codes_txt:634.81 OR discharge_icd_codes_txt:634.91 OR
discharge_icd_codes_txt:634.92 OR chief_complaint:miscarriage) AND
report_text:"emergency department"
```

140 Patients who developed disseminated intravascular coagulation in the hospital

```
(discharge_icd_codes_txt:286.6 OR report_text:"disseminated intravascular coagulation"
OR report_text:"disseminated intravascular coagulopathy") AND NOT
admit_icd_code_txt:286.6
```

141 Adult inpatients with Alzheimer's disease admitted from nursing homes with pressure ulcers

```
(report_text:"resides|resident|residing|lives|living <> **INSTITUTION" OR
report_text:"nursing home" OR report_text:"nursing facility" OR report_text:"assisted
living facility") AND (((report_text:"bed sore" OR report_text:"pressure sore" OR
report_text:"pressure ulcer") OR (discharge_icd_codes_txt:707.0* OR
discharge_icd_codes_txt:707.2*)) AND (discharge_icd_codes_txt:331.0 OR
report_text:alzheimer*))
```

142 Patients admitted with Hepatitis C and IV drug use

```
(report_text:"iv drug" OR report_text:"intravenous drug" OR report_text:heroin) AND
report_text:"admission date" AND (discharge_icd_codes_txt:070.7* OR
discharge_icd_codes_txt:070.51 OR discharge_icd_codes_txt:070.54 OR
discharge_icd_codes_txt:V02.62 OR report_text:"hepatitis c" OR report_text:"hep c")
```

143 Patients who have had a carotid endarterectomy

```
report_text:"carotid endarterectomy"
```

144 Patients with diabetes mellitus who also have thrombocytosis

```
(report_text:thrombocytosis OR discharge_icd_codes_txt:238.71) AND
(discharge_icd_codes_txt:250* OR report_text:diabetes)
```

145 Patients with lupus nephritis and thrombotic thrombocytopenic purpura

```
(((report_text:lupus OR discharge_icd_codes_txt:710.0) AND (report_text:nephritis OR
discharge_icd_codes_txt:580.* OR discharge_icd_codes_txt:582.*)) OR report_text:"lupus
nephritis") AND (report_text:"thrombotic thrombocytopenic purpura" OR report_text:ttp
OR discharge_icd_codes_txt:446.6)
```

146 Patients treated for post-partum problems including depression, hypercoagulability or cardiomyopathy

```
report_text:"post-partum" OR report_text:"post partum" OR report_text:postpartum
```

147 Patients with left lower quadrant abdominal pain

```
report_text:"left lower quadrant abdominal pain" OR report_text:"llq abdominal pain"
OR discharge_icd_codes_txt:789.04
```

148 Patients acutely treated for migraine in the emergency department

```
discharge_icd_codes_txt:346* AND report_text:"emergency department"
```

149 Patients with delirium, hypertension, and tachycardia

```
(discharge_icd_codes_txt:293.0 OR report_text:delirium) AND
(discharge_icd_codes_txt:401.* OR discharge_icd_codes_txt:405.* OR
report_text:hypertension) AND (discharge_icd_codes_txt:785.0 OR
report_text:tachycardia)
```

150 Patients who have cerebral palsy and depression

```
(report_text:depression OR discharge_icd_codes_txt:296.2* OR
discharge_icd_codes_txt:296.3* OR discharge_icd_codes_txt:296.8* OR
discharge_icd_codes_txt:311 OR discharge_icd_codes_txt:300.4 OR
discharge_icd_codes_txt:309.1) AND (report_text:"cerebral palsy" OR
discharge_icd_codes_txt:343.*)
```

151 Patients with liver disease taking SSRI antidepressants

```
(report_text:"liver disease" OR discharge_icd_codes_txt:571* OR
discharge_icd_codes_txt:572* OR discharge_icd_codes_txt:573*) AND
(report_text:"citalopram|celexa|cipramil|cipram|dalsan|recital|emocal|sepram|seropram|
citox|cital|dapoxetine|priligy|escitalopram|lexapro|cipralex|seroplex|esertia|fluoxeti
ne|depex|prozac|fontex|seromex|seronil|sarafem|ladose|motivest|flutop|fluctin|fluox|fl
uvoxamine|luvox|fevarin|faverin|dumyrox|favoxil|movox|paroxetine|paxil|seroxat|sereupi
```

```
n|aropax|deroxat|divarius|rexetin|xetanor|paroxat|loxamine|deparoc|sertraline|zoloft|l
ustral|serlain|asentra" OR report_text:ssri)
```

152 Patients with Diabetes exhibiting good Hemoglobin A1c Control (<8.0%)

```
(discharge_icd_codes_txt:250* OR report_text:diabetes) AND (report_text:"a1c 7"~4 OR
report_text:"a1c 6"~4 OR report_text:"a1c 5"~4 OR report_text:"a1c 4"~4)
```

153 Patients admitted to the hospital with end-stage chronic disease who are offered hospice care

```
report_text:hospice AND report_text:"admission date" AND NOT (report_text:trauma OR
report_text:"subdural hematoma")
```

154 Patients with Primary Open Angle Glaucoma (POAG)

```
discharge_icd_codes_txt:365.1* OR admit_icd_code_txt:365.1* OR
report_text:"open*angle" OR report_text:"primary glaucoma" OR report_text:"chronic
glaucoma" OR report_text:"intraocular pressure elevated"~5
```

155 Heart Failure (HF): Beta-Blocker Therapy for Left Ventricular Systolic Dysfunction (LVSD)

```
(discharge_icd_codes_txt:428.2*) AND
report_text:"bisoprolol|zebeta|metoprolol|lopressor|toprol|carvedilol|coreg|bucindolol
|bextra|labetalol|nadolol|pindolol|sotalol|timolol|acebutolol|atenolol|esmolol|alpreno
lol|carteolol|penbutolol|betaxolol"
```

156 Patients with depression on anti-depressant medication

```
(report_text:depression OR discharge_icd_codes_txt:296.2* OR
discharge_icd_codes_txt:296.3* OR discharge_icd_codes_txt:296.8* OR
discharge_icd_codes_txt:311 OR discharge_icd_codes_txt:300.4 OR
discharge_icd_codes_txt:309.1) AND
report_text:"abilify|ariprazole|adapin|doxepin|anafranil|clomipramine|aplenzin|bupropi
on|asendin|amoxapine|aventyl|nortriptyline|celexa|citalopram|cymbalta|duloxetine|desyr
el|trazodone|effexor|venlafaxine|emsam|selegiline|etrafon|perphenazine|amitriptyline|e
lavil|amitriptyline|endep|lexapro|escitalopram|limbitrol|chlordiazepoxide|marplan|isoc
arboxazid|nardil|phenelzine|norpramin|desipramine|oleptro|trazodone|pamelor|parnate|tr
anylcypromine|paxil|paroxetine|pexeva|paroxetine|prozac|fluoxetine|pristiq|desvenlafax
ine|remeron|mirtazapine|sarafem|seroquel|quetiapine|serzone|nefazodone|sinequan|doxepi
n|surmontil|trimipramine|symbyax|olanzapine|tofranil|imipramine|triavil|perphenazine|a
mitriptyline|viibryd|vilazodone|vivactil|protriptyline|wellbutrin|zoloft|sertraline|zy
prexa|olanzapine|fluvoxamine|anti-depressant|SSRI|TSA|MAOI"
```

157 Patients admitted to hospital with symptomatic cervical spine lesions

```
((discharge_icd_codes_txt:721.0 OR 721.1 OR 722.0 OR 722.4 OR 723.0) OR
(report_text:"cervical spine" AND (report_text:spondylosis OR report_text:osteophytes
OR report_text:"numbness weakness"~5))) AND report_text:"admission date:"
```

158 Patients with esophageal cancer who develop pericardial effusion

```
(report_text:"esophageal cancer" OR report_text:"esophagus cancer"~4 OR
discharge_icd_codes_txt:150*) AND (report_text:"pericardial effusion" OR
discharge_icd_codes_txt:423.9)
```

159 Patients with cerebral edema secondary to infection

```
(discharge_icd_codes_txt:348.5 OR report_text:"cerebral edema") AND
(report_text:infection OR report_text:infectious OR report_text:meningitis OR
report_text:cefdinir OR report_text:encephalitis OR report_text:"brain abscess" OR
report_text:acyclovir) AND NOT report_text:carcinomatous
```

160 Patients with Low Back Pain who had Imaging Studies

```
chief_complaint:"back pain" AND discharge_icd_codes_txt:724.* AND (report_text:CT OR
report_text:MRI OR report_text:radiograph OR report_text:x*ray OR
report_text:myelogram OR report_text:radiology)
```

161 Patients with adult respiratory distress syndrome

```
(discharge_icd_codes_txt:518.81 OR discharge_icd_codes_txt:518.82 OR
report_text:"adult respiratory distress syndrome" OR report_text:"acute respiratory
distress syndrome" OR report_text:ards) AND NOT(report_text:"**AGE[birth-12" OR
report_text:"**AGE[in teens")
```

162 Patients with hypertension on anti-hypertensive medication

```
(report_text:"hypertension" OR report_text:"high blood pressure" OR
discharge_icd_codes_txt:401* OR discharge_icd_codes_txt:402* OR
discharge_icd_codes_txt:403* OR discharge_icd_codes_txt:404* OR
discharge_icd_codes_txt:405*) AND
(report_text:"indapamide|chlorthalidone|metolazone|captopril|capoten|bumetanide|furose
mide|torsemide|epitazide|hydrochlorothiazide|chlorothiazide|bendroflumethiazide|amilor
ide|triamterene|spironolactone|doxazosin|phentolamine|phenoxybenzamine|prazosin|terazo
sin|tolazoline|carvedilol|labetalol|diltiazem|verapamil|eplerenone|clonidine|guanabenz
|methyldopa|moxonidine|guanethidine|reserpine" OR report_text:*olol OR
report_text:*dipine OR report_text:*opril OR report_text:*april OR report_text:*sartan
OR report_text:"beta blocker" OR report_text:"ACE inhibitor" OR report_text:"calcium
channel blocker" OR report_text: "ccb" OR report_text:"arb" OR report_text:"diuretic")
```

163 Patients treated for lower extremity chronic wound

```
report_text:"chronic|nonhealing|non-healing leg|foot|ankle|knee|thigh|calf
ulcer|ulceration|wound"~6 OR report_text:"chronic|nonhealing|non-healing lower
extremity ulcer|ulceration|wound"~6 OR report_text:"diabetic wound|ulcer" OR
((discharge_icd_codes_txt:707.06 OR discharge_icd_codes_txt:707.07 OR
discharge_icd_codes_txt:707.1*) AND report_text:"wound|ulcer") OR
((discharge_icd_codes_txt:707.8 OR discharge_icd_codes_txt:707.9) AND
report_txt:"leg|foot|ankle|knee|thigh|calf ") OR ((discharge_icd_codes_txt:440.23 OR
discharge_icd_codes_txt:459.3* OR discharge_icd_codes_txt:459.1*) AND
report_txt:"leg|foot| ankle|knee|thigh|calf")
```

164 Adults under age 60 undergoing alcohol withdrawal

```
(report_text:"alcohol withdrawal"~3 OR discharge_icd_codes_txt:291.0 OR
discharge_icd_codes_txt:291.81) AND (report_text:"**AGE[<> 20s" OR
report_text:"**AGE[<> 30s" OR report_text:"**AGE[<> 40s" OR report_text:"**AGE[<>
50s")
```

165 Patients who have gluten intolerance or celiac disease

```
(discharge_icd_codes_txt:579.0 OR report_text:gluten OR report_text:"sprue" OR
report_text:"celiac disease" OR report_text:celiacs OR report_text:"celiac
enteropathy") AND NOT (report_text:"no evidence of sprue" OR report_text:"negative for
sprue")
```

166 Patients who have hypoaldosteronism and hypokalemia

```
(discharge_icd_codes_txt:255.4 OR report_text:hypoaldosteron* OR report_text:addison
OR report_text:"congenital adrenal hyperplasia" OR report_text:"adrenal
```

```
insufficiency") AND (discharge_icd_codes_txt:276.8 OR report_text:hypokalemi* OR
report_text:hypopotass*)
```

167 Patients with AIDS who develop pancytopenia

```
(discharge_icd_codes_txt:V08 OR discharge_icd_codes_txt:042 OR (report_text:aids AND
NOT report_text:"hearing aids")) AND (report_text:pancytopenia OR
discharge_icd_codes_txt:284.1)
```

168 Patients with Coronary Artery Disease with Prior Myocardial Infarction on Beta-Blocker Therapy

```
(discharge_icd_codes_txt:414.0* OR report_text:"coronary artery disease") AND
(discharge_icd_codes_txt:410.* OR discharge_icd_codes_txt:412* OR
report_text:"myocardial infarction") AND
(report_text:"abetalol|trandate|acebutolol|sectral|bisoprolol|zebeta|esmolol|brevibloc
|propranolol|inderal|atenolol|tenormin|labetalol|normodyne|trandate|metoprolol|lopress
or|toprol|nadolol|corgard|nebivolol|bystolic|penbutolol|levatol|carvedilol|coreg|pindo
lol|sotalol|timolol|alprenolol|bucindolol|carteolol|oxprenolol|penbutolol|betaxolol|bi
soprolol|celiprolol|nebivolol" OR report_text:"beta blocker" OR report_text:"beta
blockade")
```

169 Elderly patients with subdural hematoma

```
(report_text:"subdural hematoma" OR discharge_icd_codes_txt:852.21) AND
NOT(report_text:"question of subdural hematoma") AND (report_text:"**AGE[<> 70s" OR
report_text:"**AGE[<> 80s" OR report_text:"**AGE[90+")
```

170 Adult patients who presented to the emergency room with suicide attempts by drug overdose

```
(discharge_icd_codes_txt:E950* OR ((discharge_icd_codes_txt:[960 977] OR
report_text:"drug overdose") AND (report_text:suicide OR report_text:kill)) OR
discharge_icd_codes_txt:"E980.0 |E980.1|E980.2|E980.3|E980.4|E980.5") AND
(report_text:"**AGE[<> 20s" OR report_text:"**AGE[<> 30s" OR report_text:"**AGE[<>
40s" OR report_text:"**AGE[<> 50s" OR report_text:"**AGE[<> 60s" OR
report_text:"**AGE[<> 70s" OR report_text:"**AGE[<> 80s" OR report_text:"**AGE[90+")
AND report_text:"emergency department"
```

171 Patients with thyrotoxicosis treated with beta-blockers

```
(report_text:thyrotoxicosis OR report_text:"thyroid storm" OR
discharge_icd_codes_txt:242*) AND
report_text:"abetalol|trandate|acebutolol|sectral|bisoprolol|zebeta|esmolol|brevibloc|
propranolol|inderal|atenolol|tenormin|labetalol|normodyne|trandate|metoprolol|lopresso
r|toprol|nadolol|corgard|nebivolol|bystolic|penbutolol|levatol|carvedilol|coreg|bucind
olol|bextra|pindolol|sotalol|timolol"
```

172 Patients with peripheral neuropathy and edema

```
(discharge_icd_codes_txt:782.3 OR report_text:"lower extremity edema" OR
report_text:"foot|heel|ankle edema"~4) AND NOT(report_text:"no lower extremity edema")
AND (discharge_icd_codes_txt:337.00 OR discharge_icd_codes_txt:337.09 OR
discharge_icd_codes_txt:337.1 OR discharge_icd_codes_txt:337.1 OR
discharge_icd_codes_txt:356.0 OR discharge_icd_codes_txt:356.8 OR
discharge_icd_codes_txt:356.9 OR discharge_icd_codes_txt:357 OR
report_text:"peripheral neuropathy")
```

173 Patients over 65 who had Pneumonia Vaccination Status presently or previously

```
(report_text:"pneumonia|pneumococcal vaccine|vaccination|immunization" OR
report_text:pneumovax OR report_text:ppsv OR discharge_icd_codes_txt:V03.82 OR
```

```
discharge_icd_codes_txt:V06.6) AND (report_text:"**AGE[<> 60s" OR
report_text:"**AGE[<> 70s" OR report_text:"**AGE[<> 80s" OR report_text:"**AGE[90+")
```

174 Elderly patients with ventilator-associated pneumonia

```
((report_text:"nosocomial pneumonia" AND report_text:ventilator) OR
report_text:"ventilator <> pneumonia" OR report_text:vap) AND (report_text:"**AGE[<>
70s" OR report_text:"**AGE[<> 80s" OR report_text:"**AGE[90+")
```

175 Elderly patients with endocarditis

```
(discharge_icd_codes_txt:421* OR discharge_icd_codes_txt:424.9* OR
report_text:endocarditis) AND (report_text:"**AGE[<> 70s" OR report_text:"**AGE[<>
80s" OR report_text:"**AGE[90+")
```

176 Patients with Heart Failure (HF) on Angiotensin-Converting Enzyme (ACE) Inhibitor or Angiotensin Receptor Blocker (ARB) Therapy for Left Ventricular Systolic Dysfunction (LVSD)

```
((discharge_icd_codes_txt:428.1 OR discharge_icd_codes_txt:428.2* OR
discharge_icd_codes_txt:429.9* OR report_text:"heart failure" OR report_text:chf) AND
(report_text:"left ventricular systolic dysfunction" OR report_text:"left ventricular
dysfunction")) AND
(report_text:"benazepril|lotensin|captopril|capoten|enalapril|vasotec|vasotec|fosinopr
il|monopril|lisinopril|prinivil|zestril|moexipril|univasc|perindopril|aceon|quinapril|
accupril|ramipril|altace|trandolapril|mavik|zofenopril|enalapril|imidapril" OR
report_text:"candesartan|atacand|eprosartan|teveten|iresartan|avapro|losartan|cozaar|o
lmesartan|benicar|telmisartan|micardis|valsartan|diovan|azilsartan")
```

177 Patients treated for depression after myocardial infarction

```
(discharge_icd_codes_txt:410* OR discharge_icd_codes_txt:411.0 OR
report_text:"myocardial infarction") AND
report_text:"abilify|ariprazole|adapin|doxepin|anafranil|clomipramine|aplenzin|bupropi
on|asendin|amoxapine|aventyl|nortriptyline|celexa|citalopram|cymbalta|duloxetine|desyr
el|trazodone|effexor|venlafaxine|emsam|selegiline|etrafon|perphenazine|amitriptyline|e
lavil|amitriptyline|endep|amitriptyline|lexapro|escitalopram|limbitrol|amitriptyline|c
hlordiazepoxide|marplan|isocarboxazid|nardil|phenelzine|norpramin|desipramine|oleptro|
trazodone|pamelor|nortriptyline|parnate|tranylcypromine|paxil|paroxetine|pexeva|paroxe
tine|prozac|fluoxetine|pristiq|desvenlafaxine|remeron|mirtazapine|sarafem|fluoxetine|s
eroquel|quetiapine|serzone|nefazodone|sinequan|doxepin|surmontil|trimipramine|symbyax|
fluoxetine|olanzapine|tofranil|imipramine|triavil|perphenazine|amitriptyline|viibryd|v
ilazodone|vivactil|protriptyline|wellbutrin|bupropion|zoloft|sertraline|zyprexa|olanza
pine"
```

178 Patients with metastatic breast cancer

```
((report_text:metast* OR discharge_icd_codes_tx:196* OR discharge_icd_codes_txt:197*
OR discharge_icd_codes_txt:198*) AND (discharge_icd_codes_txt:174* OR
discharge_icd_codes_txt:175* OR discharge_icd_codes_txt:v10.3)) OR
report_text:"metastatic breast cancer"
```

179 Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression

```
report_text:"amisulpride|solian|aripiprazole|abilify|asenapine|saphris|blonanserin|lon
asen|clotiapine|entumine|clozapine|clozaril|iloperidone|fanapt|lurasidone|latuda|mosap
ramine|cremin|olanzapine|zyprexa|paliperidone|invega|perospirone|lullan|quetiapine|ser
oquel|remoxipride|roxiam|risperidone|risperdal|sertindole|serdolect|sulpiride|sulpirid
|eglonyl|ziprasidone|zeldox|geodon|zotepine|nipolept" AND NOT
```

```
(discharge_icd_codes_txt:295*) AND NOT (discharge_icd_codes_txt:296*) AND NOT
report_text:"schizophrenia|bipolar"
```

180 Patients with cancer who developed hypercalcemia

```
((report_text:hypercalcemia OR discharge_icd_codes_txt:275.42) AND
(discharge_icd_codes_txt:14* OR discharge_icd_codes_txt:15* OR
discharge_icd_codes_txt:16* OR discharge_icd_codes_txt:17* OR
discharge_icd_codes_txt:18* OR discharge_icd_codes_txt:19* OR
discharge_icd_codes_txt:200.* OR discharge_icd_codes_txt:201.* OR
discharge_icd_codes_txt:202.* OR discharge_icd_codes_txt:203.* OR
discharge_icd_codes_txt:204.* OR discharge_icd_codes_txt:205.* OR
discharge_icd_codes_txt:206.* OR discharge_icd_codes_txt:207.* OR
discharge_icd_codes_txt:208.* OR discharge_icd_codes_txt:209.0* OR
discharge_icd_codes_txt:209.1 OR discharge_icd_codes_txt:209.2* OR
discharge_icd_codes_txt:209.3* OR report_text:cancer)) OR report_text:"hypercalcemia
of malignancy"
```

181 Patients being evaluated for secondary hypertension

```
(discharge_icd_codes_txt:405* OR report_text:"secondary hypertension" OR
report_text:"secondary causes of hypertension") AND NOT(report_text:"pulmonary|portal
hypertension" OR report_text:"secondary to hypertension")
```

182 Patients with Ischemic Vascular Disease

```
discharge_icd_codes_txt:410.?1 OR discharge_icd_codes_txt:411* OR
discharge_icd_codes_txt:413* OR discharge_icd_codes_txt:414.0* OR
discharge_icd_codes_txt:414.8 OR discharge_icd_codes_txt:414.9 OR
discharge_icd_codes_txt:429.2 OR discharge_icd_codes_txt:433* OR
discharge_icd_codes_txt:434* OR discharge_icd_codes_txt:440.1 OR
discharge_icd_codes_txt:440.2* OR discharge_icd_codes_txt:440.4 OR
discharge_icd_codes_txt:444* OR discharge_icd_codes_txt:445* OR report_text:"ischemic
vascular|cardiovascular|cerebrovascular|heart" OR report_text:"ischemic peripheral
artery"
```

183 Patients presenting to the emergency room with acute vision loss

```
(chief_complaint:blur* OR chief_complaint:"vision loss" OR report_text:"vision loss"
OR report_text:"visual disturbance|impairment" OR discharge_icd_codes_txt:368.40 OR
discharge_icd_codes_txt:377.30 OR discharge_icd_codes_txt:438.7 OR
discharge_icd_codes_txt:368.9) AND (report_text:"emergency department" OR
report_text:"emergency room")
```

184 Patients with Colon Cancer who had Chemotherapy

```
(report_text:"colon cancer" OR report_text:"colorectal cancer" OR
discharge_icd_codes_txt:153* OR discharge_icd_codes_txt:V10.05) AND
report_text:"chemotherapy|fluorouracil|5-
FU|leucovorin|xeloda|camptosar|eloxatin|avastin|erbitux|vectibix|camptosar|eloxatin|av
astin|folfox|folfiri"
```

185 Patients who develop thrombocytopenia in pregnancy

```
(discharge_icd_codes_txt:v22 OR discharge_icd_codes_txt:v23 OR
report_text:"pregnant|pregnancy|hellp") AND (discharge_icd_codes_txt:287.3 OR
discharge_icd_codes_txt:287.4 OR discharge_icd_codes_txt:649.3 OR
report_text:thrombocytopen*)
```